Assessing the limits on size-pitch mapping reveals the interplay between top-down and bottom-up influences on crossmodal correspondences

Allegra Indraccolo[*], Claudia Del Gatto[*], Tiziana Pedale[+], Valerio Santangelo[^,+], Charles Spence[°], & Riccardo Brunetti[*]

* Experimental and Applied Psychology Laboratory, Department of Human Sciences, Università Europea di Roma, Italy

^ Department of Philosophy, Social Sciences & Education, University of Perugia, Italy

+ Functional Neuroimaging Laboratory, IRCCS Santa Lucia, Rome, Italy

° Department of Experimental Psychology, Oxford University, UK

Abstract

Certain sensory dimensions, such as visual size and auditory pitch, are consistently associated, resulting in performance facilitation or inhibition. The mechanisms underlying these crossmodal correspondences are still the subject of debate: The relative or absolute nature of crossmodal mappings is connected to this debate, as an absolute mapping may indicate a bottom-up process, whereas a relative one may be evidence of stronger top-down influences. Three experiments were conducted (including overall N = 207 participants), based on two different tasks, designed to explore a wide range of size-pitch crossmodal mappings. In Experiment 1, the participants were instructed to freely manipulate stimuli varing along a given dimension to 'match' the other. The results revealed evidence for a quasi-absolute mapping, but the correspondences shifted depending on the participants' auditory or visual attentional focus. In Experiment 2, the participants performed a visual speeded categorization task, involving a wide range of auditory task-irrelevant pitches, including the "preferred" ones, estimated on the basis of the results of Experiment 1. The results revealed a rather relative mapping, corroborating a top-down influence on the correspondence effect. Experiment 3 was designed to determine whether the relative mapping involved has boundary. The results confirmed that the larger the interval between pitches (i.e., more perceptually salient), the stronger the congruence effect, thus highlighting bottom-up facilitation. Taken together, these findings reveal that the size-pitch correspondences are sensitive to task-related top-down factors, as well as to stimulus-related bottom-up influences, ultimately revealing the adaptive nature of this kind of multisensory integration.

Abstract word count: 247

Keywords: Multisensory integration; Attention; Contextual effects; Sensory association range

Public significance statements

- This study provided evidence that there is no pre-determined audiovisual mapping of stimuli in the association between visual size and auditory pitch: A solid crossmodal congruency effect can be elicited in a seemingly completely relative manner, though its magnitude may vary according to the specific stimuli and task used.

- The study highlights how attention plays a crucial role in multisensory integration, even when dealing with crossmodal correspondences (e.g., seemingly automatic associations between specific sensory dimensions). The experiments specifically reveal significant effects of both top-down (task-related) and bottom-up (stimuli-related) influences, showing the interplay between these different kinds of information processing.

Introduction

A considerable body of empirical research in the field of multisensory perception has recently been dedicated to the study of crossmodal correspondences: these are the consensual tendencies people show to associate different stimuli or features across sensory dimensions (see Spence, 2011; Walker, 2016, for reviews). While this line of empirical research has by now uncovered a wide array of different cross-sensory couplings (e.g., Deroy, Crisinel, & Spence, 2013), one of the most frequently-studied perceptual features that has been found to be associated with other sensory dimensions is auditory pitch (Spence, 2019a, for a review). Auditory pitch has been shown to correspond crossmodally with several other features, most notably in the visual domain, with size (Evans & Treisman, 2010; Parise & Spence, 2009, 2012), elevation (Ben-Artzi & Marks, 1995; Chiou & Rich, 2012), brightness (Klapetek, Ngo, & Spence, 2012), hue (Melara, 1989; Melara & O'Brien, 1987), shape (O'Boyle & Tarte, 1980; Parise & Spence, 2012). These correspondences have been revealed and investigated using a wide range of different behavioural tasks, from speeded classification (see Marks, 2004, for a review) through to implicit associations tasks (Parise & Spence, 2012), and from exogenous crossmodal spatial attentional cuing studies (Chiou & Rich, 2012, 2015) to working memory paradigms (Brunetti, Indraccolo, Mastroberardino, Spence, & Santangelo, 2017). Certain of these correspondences, such as the size-pitch correspondence, have been confirmed in other mammals (in primates: Kalan, Mundry, & Boesch, 2015; in dogs: Taylor, Reby, & McComb, 2011), in reptiles (Loconsole, Stancher, & Versace, 2023), and have been recently found to be related to signature activity in low-level sensory cortices (Sciortino & Kayser, 2023).

While crossmodal correspondences involving auditory pitch were initially attributed to automatic, bottom-up processes (e.g., similar to synaesthesia, based on the systematic induction of sensory attributes; see Martino & Marks, 2001), more recently the influence of

various top-down factors has gradually been revealed (Getz & Kubovy, 2018; Klapetek *et al.*, 2012). The ongoing debate between bottom-up and top-down factors is therefore sustained by sets of different findings pointing towards more automatic, bottom-up processing (see Parise & Spence, 2012), while others revealed stronger top-down influences (e.g., goal-oriented; Jamal, Lacey, Nygaard, & Sathian, 2017; McCormick, Lacey, Stilla, Nygaard, & Sathian, 2018). For example, using the Implicit Association Test (IAT), Parise and Spence (2012) reported audio-visual correspondences, which occurred at the response selection level, and could be an outcome of an automatic process. In five experiments, the authors demonstrated that certain crossmodal correspondences (e.g. visual size-pitch correspondence) were automatically encoded, modulating the participants' RTs even in the fastest responses (those occurring in under 400ms), thus providing evidence for the role of bottom-up processes (Parise & Spence, 2012). On the other hand, Klapetek and colleagues (2012) performed experiments using the "pip-and-pop effect", and demonstrated a crossmodal facilitation of to-be-searched for visual targets when they were paired (synchronized) with task-irrelevant sounds. Their results revealed that crossmodal facilitation in this experimental paradigm was crucially mediated by top-down attention, as facilitatory effects appeared only when participants were explicitly informed about that specific crossmodal congruence.

In general, attention has been shown to have top-down modulatory influence on multisensory effects, even when they are strong and well-established (such as the McGurk or the audiovisual ventriloquism effect: see Alsius, Navarra, Campbell, & Soto-Faraco, 2005; Duyar, Pavan, & Kafaligonul, 2022; for a review, see Santangelo & Macaluso, 2012; see also Del Gatto, Brunetti, & Delogu, 2016, for evidence from memory-related tasks), although contradictory evidence does exists. For instance, Evans (2020) has argued that crossmodal interactions between corresponding basic features are independent of selective attention (and so less susceptible to top-down influences). Across three experiments, Evans systematically

altered the discriminability of stimuli in speeded categorization and visual search tasks, without obtaining a significant reduction of the crossmodal congruence effects between visual elevation and pitch. However, it is important to note that these results were obtained with "indirect" tasks (Evans & Treisman, 2010), in which attention was focused on different auditory and visual features from those that were expected to show a crossmodal correspondence.

Additional proof that top-down influences might have an impact on crossmodal effects comes from the asymmetrical effect, related to the specific modality or task used, that have been observed in certain studies: visual stimuli appear to affect other modalities more than vice versa (Spence, 2019b), which is referred to as human visual dominance (Hutmacher, 2019; Posner, Nissen, & Klein, 1976). These asymmetrical effects can sometimes be reversed in favour of other sensory modalities (e.g., audition), according to the specific task used (e.g., temporal judgments; see Varghese, et al., 2017). This is in accord with the *modality appropriateness hypothesis* (Welch & Warren, 1980), stating that in the case of bimodal sensory stimulation, the sense with the higher acuity with respect to the most critical aspects of a task plays a predominant role on how multisensory inputs are integrated.[1] This may well apply to crossmodal correspondences, that are commonly considered to be bi-directional (e.g., with mutually affecting sensory modalities, Deroy & Spence, 2013; Walker, 2016). Spence and Deroy (2013) proposed alternatively that different results may depend on the fact that crossmodal mappings can be generated by multiple mechanisms, or be of different kinds (e.g., statistical, structural, and semantic; Spence, 2011; Spence and Deory, 2013). Indeed, taking into account the classification proposed by Spence (2011), some kind of crossmodal correspondences, such as that between auditory pitch and visual size, can be formed based on the correlation between certain dimensions in the world around us. In the

[1] It is worth noting how the *modality appropriateness hypothesis* has been superseded by approaches using *maximum likelihood estimation* (Alais, & Burr, 2019; Andersen, Tiippana, & Sams, 2005).

case of these "statistical correspondences," such as pitch/size, it makes sense that our brains take advantage of the regularities existing in the natural environment when deciding which of the many possible stimuli to integrate, thus revealing different dependencies with respect to bottom-up and top-down perceptual/attentional processes (cf. Spence & Di Stefano, 2022).

Talsma, Senkowski, Soto-Faraco, and Woldorff (2010) proposed a framework to understand the multifaceted involvement of attention in multisensory integration: when stimulus salience is low, multisensory integration may happen more or less pre-attentively; when sensory salience is increased, bottom-up processes can act as a facilitator for attentional selection and discrimination; finally, when multiple stimuli compete for processing resources, top-down selective attention is likely to be necessary in order for multisensory integration processes to take place.

Another relevant aspect of this debate concerns the absolute or relative nature of these couplings (see Brunetti, Indraccolo, Del Gatto, Spence, & Santangelo, 2018): namely, whether a specific correspondence can produce effects only with a pre-determined audiovisual mapping of stimuli (e.g., similar to what happens in the case of synaesthesia, where specific values of a dimension are naturally paired by some individuals with specific values of a different sensory dimension; Deroy & Spence, 2013) or else these correspondences are based on contingencies and context (e.g., they can arise based on specific relative relationships between different sensory dimensions). While some authors have apparently obtained evidence of absolute mappings between certain dimensions (see Guzman-Martinez, Ortega, Grabowecki, Mossbridge, & Suzuki, 2012, for auditory amplitude modulation and visual spatial frequency), the vast majority of results have highlighted the existence of relative mappings (see Spence, 2011, for a review). Brunetti et al. (2018), among others, conducted a study to measure correspondences between auditory pitch and visual size, systematically manipulating the sequence of stimuli and demonstrating that a congruence

effect can arise without any apparent unconstrained mapping: the congruence can be based on the immediate previous events and can thus adjust to the context in which stimulation takes place. This and other results (Chiou & Rich, 2015; see also Spence, 2019a) appear to point in the direction of a relative mapping between different dimensions, especially when that mapping includes pitches among the dimensions involved. Moreover, according to Spence and Di Stefano (2022), the structure of the involved dimensions themselves (e.g., whether they are prothetic or metathetic; see Stevens, 1957) may constrain, or modulate, these correspondences (i.e., in terms of which of the many possible crossmodal correspondences may come to mind most naturally, or spontaneously).

It is still a matter of debate if this relative mapping is unconstrained (that is, can any value be crossmodally paired with any other?) or if some absolute restriction applies (e.g., there are boundary conditions that limit the range of possible crossmodal mappings). This matter is specifically relevant to shedding light on the nature of some (or possibly all) crossmodal correspondences. Moreover, it might give us some insight into "how crossmodal correspondences are acquired in the first place, whether they differ in terms of their perceptual consequences, and/or their resistance to retraining" (Spence, 2019a, p. 242). This issue is specifically linked to top-down attentional influences, as the less strict these limits become, namely making the mappings more adaptive and variable, the less grounds we have to interpret the effect as bottom-up, hard-wired in the perceptual system.

In the case of pitch-based crossmodal correspondences, the issue of possible constraints to relative mappings is hard to define since the sound frequencies used in each study have varied considerably and are usually only used in pairs (e.g., only two sound frequencies, one relatively low and the other relatively high, in each experiment; much more rarely three, see for an example Brunetti et al., 2018). Moreover, studies usually do not report any data on pre-testing of visual-auditory couplings (Spence, 2019a, p. 242) to maximize the

effects. The literature thus lacks specific investigations concerning the limits of these pairings. And indeed, intriguing research by Parise, Knorre, and Ernst (2014), looking specifically at the case of the pitch/visual elevation correspondence has shown that it may operate over a relatively restricted range of frequencies.

Following Spence's (2019a, p. 242) suggestions, an experiment was designed aiming at defining the nature of the spontaneous couplings between visual size and auditory pitch. By using the method of adjustment (as already applied to other crossmodal correspondences: Orchard-Mills, Van der Burg, & Alais, 2013; see also Cardozo, 1965) it is possible to define the ranges of the couplings that emerge when participants are explicitly asked to map one dimension to the other. In this task (Experiment 1), the participants were instructed to focus their attention on both dimensions, since they were both task-relevant. The results allowed us to define whether the mappings between visual size and auditory pitch show regularity across participants (namely, different participants match similar pitch heights with visual sizes and vice versa). Moreover, by asking the participants to adjust pitch according to visual size and the opposite, it was possible to assess whether the effect was symmetrical (e.g., the same pitch is coupled with the same size, and vice versa).

The results of our first study were then used in a second experiment to manipulate participants' performance in a speeded visual classification task. By systematically varying the pitches of task-irrelevant sounds that were paired with fixed visual sizes, we assessed whether the ranges revealed in Experiment 1 were, in fact, relevant (e.g., whether they maximized congruency effects), when attention was focused on only one sensory modality. If the results of Experiment 2 (where attention was focused on a single modality) were to reveal a crossmodal facilitatory effect of the specific mappings emerging from Experiment 1 (where attention was focused on both sensory modalities), it would be possible to conclude that attention plays a limited role in the size-pitch correspondence. Conversely, if the specific

mappings that were to be obtained in the first experiment failed to affect performance in the second experiment, this might be taken to suggest that attention (and then, top-down modulation) plays a crucial role in this correspondence. Moreover, by using multiple pitches coupled with the same visual stimuli, the results of the second experiment can hopefully help to clarify whether mappings continue to have facilitatory effects using a specific range of pitches variation. Finally, the audiovisual mapping effects documented in Experiment 2 were further investigated in a third experiment using a wider range of pitch intervals coupled with only two visual sizes, to additionally test size-pitch crossmodal correspondences.

EXPERIMENT 1

Methods

Participants

One-hundred undergraduates took part in the experiment for course credit (31 males; mean age = 20.4 years; SD = 1.31; range = 19–25 years) and were randomly divided into two groups of 50 participants each. A post-hoc power analysis was carried out using G*Power 3.1.9.2 (F-tests, repeated-measures ANOVA). This analysis revealed that the current sample size guaranteed a statistical power of >99% as computed by considering an effect size of 0.338, 7 measurements (corresponding to the 7 levels of the conducted ANOVA, see below the Results section) and 2 groups, and a significance level of 0.05. All the participants reported normal or corrected-to-normal hearing and vision. The participants were naïve as to the purpose of the study, which had been approved by the Ethics Review Board of the Università Europea di Roma.

Apparatus and materials

Stimulus presentation, conditions, randomization, and the recording of responses were all controlled by a custom-made script in Max (v. 7, Cycling'74) programming environment, running on a 15" 2.4 GHz MacBook Pro laptop computer. The sounds were presented through Edirol MA-15D speakers at a comfortable listening level. In order to ensure the co-localization of the auditory and visual stimuli, the display was set exactly in the middle of the two loudspeakers. Finally, all the different pitches were adjusted and measured with a sound meter to obtain a similar perceived loudness according to an equal loudness contour (Scharf, 1978).

In the Pitch adjustment task, the visual stimuli consisted of 7 black discs of increasing size: 1.2°, 2.4°, 3.6°, 5°, 6.3°, 8°, and 12° visual angles. The discs were presented at the centre of a grey (50%) display for 400 msec, flashing with a 1000 msec Inter-Onset Interval (IOI). The auditory component of the Audio adjustment task consisted of a series of sinusoidal tones with 400 msec duration (including two 5 msec ramps) presented simultaneously with the discs and adjustable in discrete whole-tone steps from a minimum of 16.3 Hz (corresponding to a C0 note) to a maximum of 7458.6 Hz (corresponding to a Bb8 note), giving rise to a total of 52 tones. A whole-tone scale was chosen over a diatonic scale in order to avoid choices based on consonance and to have the participants focus only on pitch height.

In the Size adjustment task, the auditory stimuli consisted of 7 sinusoidal tones of increasing pitch: 300, 415, 517, 738, 1025, 1380, and 1770 Hz. Each tone had a 400 msec duration (including two 5 msec ramps) and was presented intermittently with 1000 msec IOI. The visual component of the Video adjustment task consisted of a series of black discs, presented on a grey (50%) display for 400 msec, presented simultaneously with the tones.

The discs could be adjusted incrementally (of equal magnitude) in size, from 0.9° to 11.8° visual angles, giving rise to a total of 25 different sizes.

Procedure

The participants were seated in a testing room, facing a monitor placed approximately 60 cm in front of them. They were instructed to stay still in front of the screen and they received instructions. After having completed a few practice trials, the experiment began.

The participants were randomly assigned to the two adjustment tasks. In the "Pitch adjustment" group, the participants were instructed to adjust the intermittent tone to the flashing disc. They were instructed that there were no right or wrong answers in the task, and they were invited to carefully select a pitch that would harmonically match with the flashing disc, according to their own judgment[2]. After a brief training aimed at familiarizing them with the task, they were presented with a random sequence of the 7 discs of varying size. Every trial featured a flashing disc (400 msec exposure, with a 1000 msec IOI) simultaneously presented with a sinusoidal tone (400 msec long), randomly selected out of the 52 possible tones. In each trial, the participants adjusted the pitch of the tone using the up and down arrows on the keyboard until they had reached a satisfactory match between the disc and the sound. At this point, pressing the space bar would move onto the next trial. Each size was matched with a pitch 10 times, in a completely randomized order.

In the "Size adjustment" group, the participants were asked to adjust the sizes of the visually-presented discs to a series of 7 intermittent tones. They were instructed that there were no right or wrong answers in the task. They were invited to carefully select a discs size that would match with the auditory pitch, according to their own judgment. After instructions and training, auditory tones were presented intermittently at the same rate described above

[2] It is worth noting that no hint was given to participants about the existence of a possible mapping between visual dimension and auditory pitch.

and were coupled with an adjustable flashing disc. The disc presented initially in each trial was randomly selected among the 25 sizes available and the participants were able to adjust it using the keyboard exactly as described above. The 7 tones were matched with a visual size in a random order 10 times each.

The script also calculated the average pitch selected by each participant in the Pitch adjustment group and the average size of the disc selected by each participant in the Size adjustment group, for each of the 7 stimuli used in each task. These averages were calculated across the 10 repetitions of each stimulus and were used as a dependent variable for the analysis.

Both tasks included a total of 70 trials and lasted on average approximately 20 minutes.

Results

Data from the two groups ("Pitch adjustment" and "Size adjustment") were first analyzed separately. A repeated measures ANOVA was carried out on the data obtained from the "Pitch adjustment" group with Visual Size as a single independent factor with 7 levels, and the pitch values (expressed in Hz) selected by participants as a dependent factor. In case of violation of the sphericity assumption (Mauchly's sphericity test; $p < 0.05$), Greenhouse-Geisser corrected degrees of freedom were reported. Moreover, we also carried out Bayesian repeated measure ANOVA to report the strength of evidence of each main and post-hoc effect, according to the following interpretation: BF10 < 1: no evidence, 1–3: anecdotal evidence, 3–10: substantial evidence, 10–30: strong evidence, 30–100: very strong evidence, and > 100: decisive evidence (Keysers et al., 2020). The analysis revealed a significant effect of Visual Size on the chosen pitches [F (2.43, 119.22) = 23.4, $p < .001$; $\eta^2_p = 0.323$, BF10 >

100], with participants selecting progressively higher frequency pitches for smaller visual sizes (Fig. 1). A post-hoc analysis (Holm-Bonferroni corrected) revealed that the average pitch chosen for 1.2 visual degrees was significantly higher from all other pitches (all *ps* < .001; all BF10 > 81.1). The average pitches for 2.4 and 3.6 degrees were not significantly different from each other, but the average pitches chosen for 2.4 were significantly higher than the pitches chosen for all the other bigger visual degrees (all *ps* < .003; all BF10 > 95.6). The pitches chosen for the 3.6 degrees were higher than the pitches chosen for all the other larger visual degrees. However only the difference between 3.6 and 8.0 visual degrees was significant, p = .03, while the Bayesian post-hoc revealed evidence in all comparisons: 3.6 vs. 5, BF10 = 3.7; 3.6 vs. 6.3, BF10 = 14.3; 3.6 vs. 8, BF10 >100; and 3.6 vs. 12, BF10 = 5.7. All the other selected pitches for 5, 6.3, 8, and 12 visual degrees were not significantly different from each other (all *ps* = 1), even though the Bayesian post-hoc revealed substantial evidence for the difference between 5 and 8 visual degrees (BF10 = 6.3).

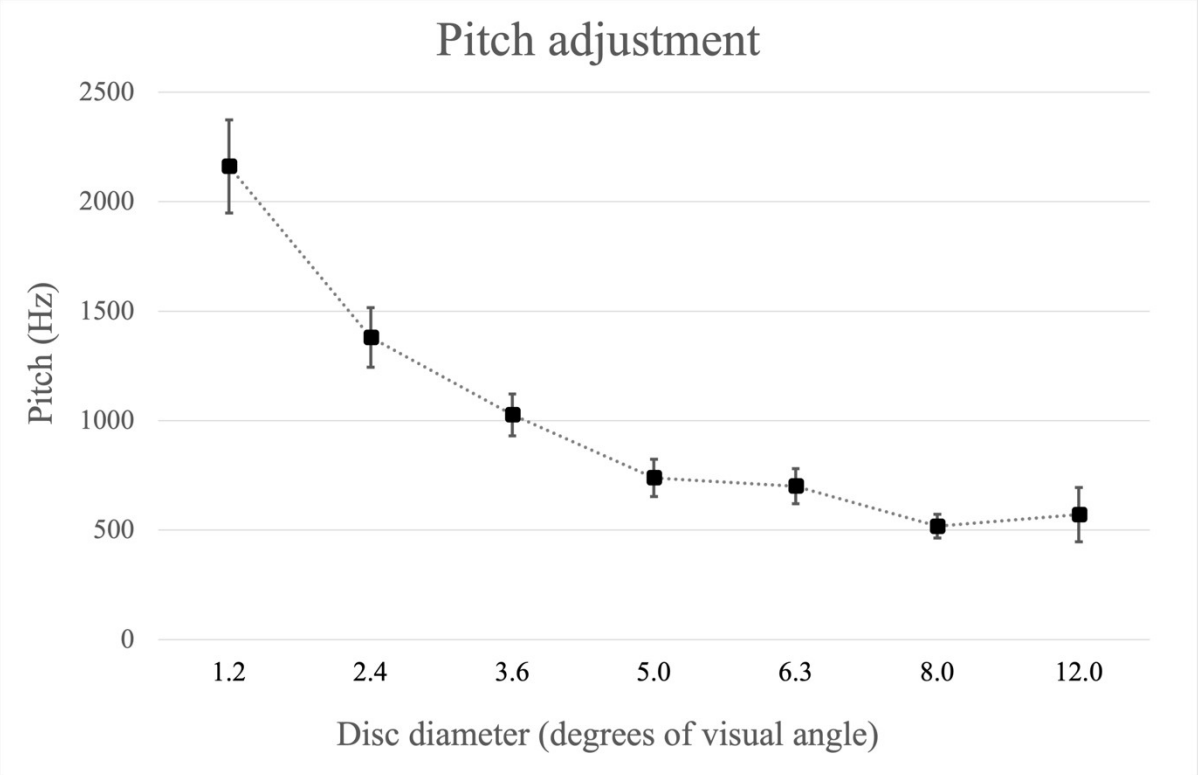


Figure 1. Results from the “Pitch adjustment” group of Experiment 1. Error bars indicate the standard error of the means.

A second repeated measures ANOVA was carried out on the data obtained from the “Size adjustment” group with Auditory Pitch as a single independent factor with 7 levels, and the size values (expressed in degrees of visual angle) selected by participants as a dependent factor. The analysis revealed a significant effect of Pitch on the chosen sizes [F (1.54, 75.33) = 25.0, $p < .001$; $\eta^2_p = 0.338$; BF10 > 100], with participants selecting progressively smaller sizes for higher pitches (Fig. 2). A post-hoc analysis (Holm-Bonferroni corrected) revealed that the average size chosen for 300Hz was significantly different from all other sizes (all *ps* < .001; all BF10 > 100). The average sizes for 415Hz and 517Hz were not significantly different from each other, but the average size chosen for 415Hz was significantly larger than the sizes chosen for all the other higher pitches (all *ps* < .022; all BF10 > 7.9). The average sizes for 517Hz were larger than the sizes selected for the higher pitches, with significant differences with the sizes selected at 1025Hz, 1380Hz, and 1770Hz (all *ps* < .006, all BF >100). Moreover, Bayesian post-hoc analyses revealed strong evidence for different sizes selected with 517Hz vs. 738Hz (BF10 = 27.7). The average size for 738Hz did not appear statistically different from the size for 1025Hz (although the BF10 was 10.55), but was significantly different from those for 1380Hz and 1770Hz (all *ps* < .025; all BF10 > 100). Finally, the selected sizes for 1025Hz, 1380Hz, and 1770Hz were not significantly different from each other (all *ps* > .24), although the Bayesian post-hoc revealed substantial evidence for the sizes chosen for 1025Hz vs. 1380Hz (BF10 > 100), and for the sizes chosen at 1025Hz vs. 1770Hz (BF10 = 8.8).

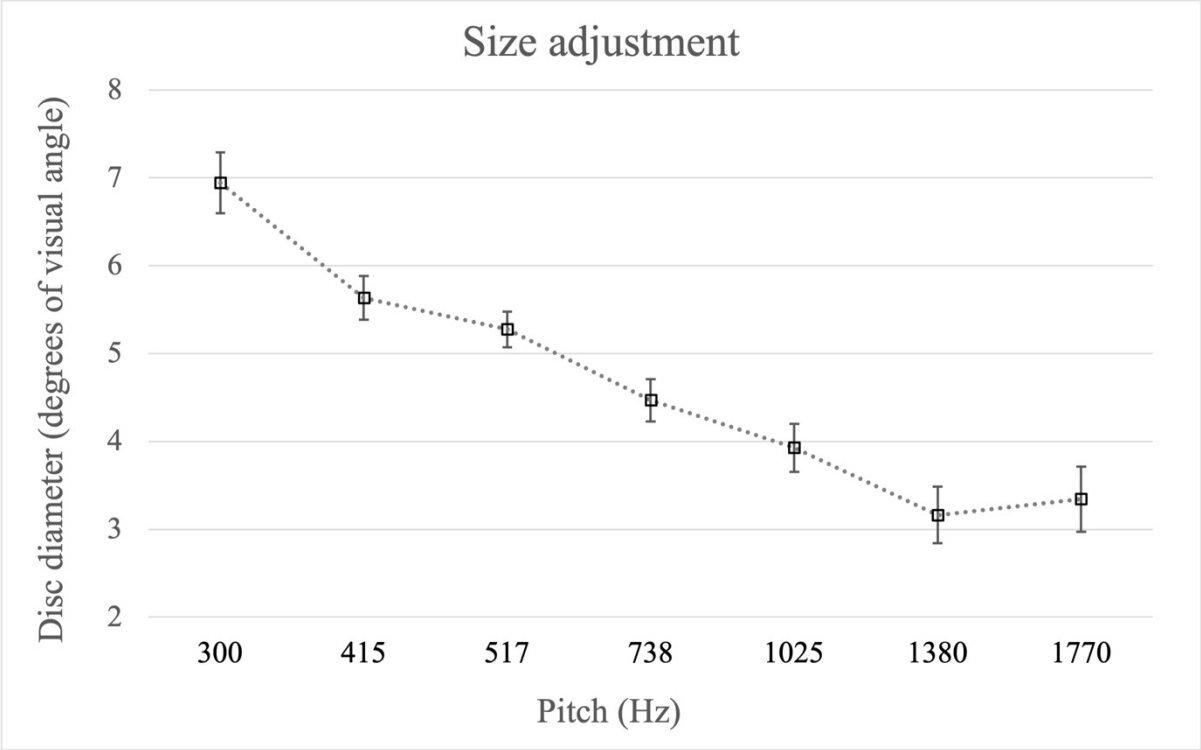


Figure 2. Results from the “Size adjustment” group in Experiment 1. Error bars indicate the standard error of the means.

Finally, the question of whether the crossmodal correspondence between visual size and auditory pitch differed between the “Pitch adjustment” and “Size adjustment” groups was evaluated. The average of the 14 audiovisual crossmodal correspondences obtained from the two groups according to each stimulus used as an input (i.e., the 7 increasing sizes of visual stimuli used as input for the pitch adjustment task, and the 7 increasing auditory pitches used as the input for the size adjustment task) were entered into a single analysis. After preparing the 14 observations, the relationships between visual size and auditory pitch in the two experiments were tested for significant differences. First, the relation between the visual size and auditory pitch relation was tested for linearity, by means of the rainbow test (Utts, 1982). The analyses yielded significant results ($p = .017$), indicating a non-linear relation. To handle this non-linear relationship, the data were log-transformed (Benoit, 2011), meaning that the

data met the criteria of linearity (the Rainbow test on the log-transformed data was not significant, p = .460), and a linear regression model was performed. The R^2 of the linear regression model with log-transformed data was 0.962, while the R^2 of the linear regression model with the original data was 0.724, thus indicating that the correspondence between audiovisual features was better explained by a logarithmic relation, than a linear relation. This was also confirmed by the AIC score since the lowest AIC score was obtained whit the model of log-transformed data (-12.67), while the AIC of the model with the original data was 207.23. Thus, the model with the log-transformed data was used to test whether the relationships between the percentage change in visual size and the percentage change in auditory pitch in the two experiments (i.e., the steepness of the solid and dotted lines in Figure 3, for the size adjustment group and the pitch adjustment group, respectively) were significantly different. The model revealed that the percentage change in auditory pitches was significantly associated with the percentage change in visual stimuli size in both groups [$F(1,10) = 176.45$, $p < .001$)]. Moreover, the model also revealed that there was a significant difference between the two groups [$F(1,10) = 6.38$, $p = .03$)], as well as a significant difference in the steepness of the regression lines of the two groups [$F(1,10) = 66.95$, $p < .001$)]. This suggests a different crossmodal correspondence depending on the role of the stimuli (fixed vs. adjustable) involved in the adjustment tasks. The results were further supported by the Bayesian analyses, which revealed the following Bayes factors: BF10 = 80 for the association between percentage change in auditory pitches and visual sizes in both groups; and a BF10 > 100 in support of a different steepness of the regression lines. There was no evidence in support of an overall difference between the “Size adjustment” and “Pitch adjustment” groups (BF10 = 0.51).

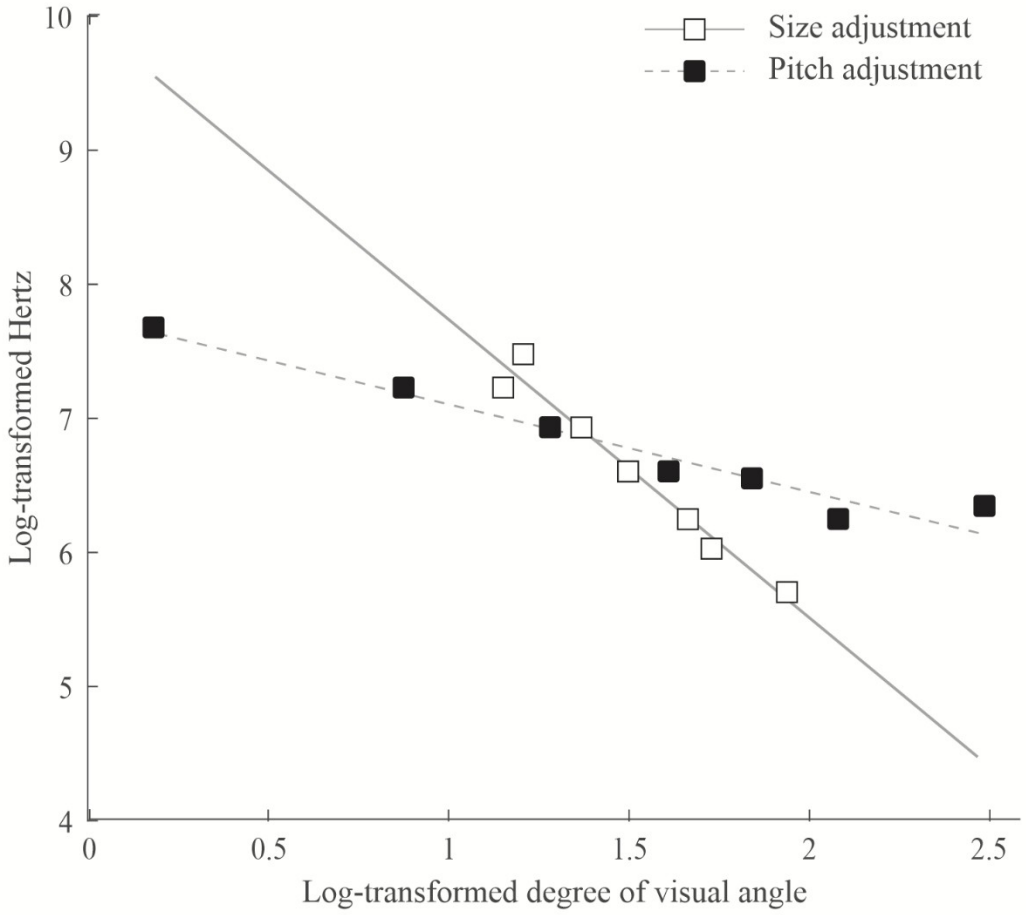


Figure 3. Interaction plot displaying the linear relation between the percentage changes in the auditory pitch (log-transformed Hz values) and the percentage changes in the visual size (log-transformed degrees of visual angle) in the “Size adjustment” and “Pitch adjustment” groups.

Discussion

The results of Experiment 1 confirmed a clear relationship between visual size and auditory pitch in a crossmodal adjustment task (Orchard-Mills et al., 2013). There are at least two features of these results that point towards an absolute relationship of the crossmodal matching couplings between pitch and size dimensions. First, the differences between at least

some of the selected pairs are significant: this shows that the stimulus pairs have been selected with consistency across participants, systematically preferring certain pitch heights or visual sizes in the continuum provided. Second, the choices did not span across the continua evenly, but logarithmically, showing that the preferences of stimuli dimensions (both pitches and sizes) did not appear to adjust fully to the boundaries of the ranges of the specific stimuli provided (e.g., the specific stimulus set used). Specifically, in the "Pitch adjustment" task, the participants matched similar pitches to the variation between 5 and 12 degrees of visual angle, while in the "Size adjustment" task, the visual sizes coupled for pitches of 1380 and 1770Hz were similar. These two features of the results would appear to point towards an emerging absolute (e.g., bottom-up) coupling. This is good agreement to what was found by Guzman-Martinez and colleagues (2012) for auditory amplitude modulation and visual spatial frequency, using a similar task.

However, other aspects of the results point in a completely different direction. If crossmodal mappings between visual size and auditory pitch were to be determined in an entirely bottom-up manner, it would be expected that the relationship should be similar regardless of the specific adjustment task (e.g., adjusting pitch to size or size to pitch): in other words, a symmetrical effect, as has often been documented in research on the crossmodal correspondences (Deroy & Spence, 2013; Walker, 2016). In these analyses, a direct comparison of the slopes of the psychometric functions obtained shows that the relationships obtained do not match. This significant mismatch may be generated by the different attentional focus necessary in the two tasks: while it is true that in this task both dimensions are task-relevant (e.g., matching one dimension to the other means that participants must consider both dimensions at all times), it is true that the fixed stimulus and the adjustable one might have recruited different amounts of attention. Depending on which of the two dimensions is adjustable (visual size or auditory pitch), the participants might have

set/adjusted their their attentional focus differently (for other asymmetrical effects in crossmodal pairings, see Spence, 2019b).

The participants' choices thus seem to change according to the task at hand, possibly showing an attention-dependent top-down effect on crossmodal couplings (Klapetek et al., 2012). This points towards the possibility that the specific task may affect the specific mappings, generating the illusion (when observed individually) of a quasi-absolute mapping, whereas other tasks may reveal more of a relative mapping (as expected on the basis of an overwhelming body of research: Brunetti et al., 2018; Chiou & Rich, 2015; Spence, 2011, 2019a). On the other hand, as Mellers and Birnbaum (1982) highlighted more than four decades ago, the stimulus distribution, namely, the specific set of stimuli that are used in a task, may affect judgments. Specifically, the perceiver might compare the relative position of a stimulus within its contextual distribution, with the relative position of a stimulus of another modality to its own contextual distribution. Along this line, any change in the stimulus distribution in a specific task may be expected to affect the correspondences between different stimulus dimensions (see also the well-known biases in psychophysics related to stimulus range and response frequency equalization, e.g., Erlebacher & Sekuler, 1971; Sekuler & Erlebacher, 1971). However, in our study the participants did not have a specific set of stimuli, but they could flexibily adjust, in a continuous way, the specific dimensions: the large ranges available for the adjustable dimensions makes it unlikely for the difference between the slopes to depend on the specific set of the stimulus distribution used.

In order to investigate whether the specific adjustment task used may have affected the results that were obtained, a second experiment was designed. To check if the specific crossmodal correspondences identified by participants between size and pitch chosen by the participants in Experiment 1 are somehow more effective in generating a crossmodal congruence effect, these results were used as a pre-test of crossmodal pairings (see Spence,

2019a) for Experiment 2. A speeded categorization task was created, using the preferred couplings of Experiment 1 strategically to verify whether these elicited greater crossmodal congruence effects when directly compared to other possible couplings in a new and independent (from the estimation of Experiment 1) sample of participants. To further clarify the possible role of top-down modulation on crossmodal correspondences, in Experiment 2, the participants were instructed to focus their attention on the visual dimension, while the sounds were task-irrelevant.

EXPERIMENT 2

Method

Participants

Forty undergraduates took part in the experiment for course credit (15 males; mean age = 21.7 years; SD = 3.06; range = 19–30 years). We checked whether the sample size was appropriate to reach sufficient statistical power. A post-hoc power analysis was carried out using G*Power 3.1.9.2 (F-tests, repeated-measures ANOVA). This analysis revealed that the current sample size guarantees a statistical power of 99% as computed by considering an effect size of 0.32, 7 measurements (corresponding to the 7 levels of the conducted ANOVA, see below the Results section), and a significance level of 0.05. All the participants reported normal or corrected-to-normal hearing and vision. The participants were naïve as to the purpose of the study, which has been approved by the Ethics Review Board of the Università Europea di Roma.

Apparatus and materials

Stimulus presentation, conditions, pseudo-randomization, and the recording of responses were all controlled by a custom-made script in Max (v. 8, Cycling'74) programming environment, running on a 2.4 GHz Macmini Desktop computer. The visual targets for the speeded classification task consisted of three black discs of respectively 1.2° (small - S), 3.6° (medium - M), and 8.0° (large - L) visual angle diameter; the discs were presented individually at the centre of a grey (50%) 21" display for 300 msec. In order to ensure the co-localization of the auditory and visual stimuli, the display was set exactly in the middle of the two loudspeakers (Edirol MA-15D), which were used to present task-irrelevant sounds at a comfortable listening level. The auditory stimuli consisted of nine sinusoidal tones, roughly corresponding to musical notes: 415Hz (Ab4), 517Hz (~C5), 660Hz (E5), 830Hz (Ab5), 1026Hz (~C6), 1318Hz (E6), 1661Hz (Ab6), 2161Hz (~C7), 2637Hz (E7). According to the results of Experiment 1, three pitches were selected matched, on average, to the 3 selected visual sizes, and each of them was altered of approximately a major third (4 semitones) either upwards or downwards, generating the final 9 pitches. This selection repeated approximately the same three notes in order to avoid the perception of the sequence as a tonal musical melody. All the different pitches were adjusted and measured with a sound meter to obtain a similar perceived loudness according to an equal loudness contour (Scharf, 1978). Each tone was presented for 300 msec and included two 5 msec fading-ramps.

Procedure and design

Administration conditions (testing room, distance from the screen) were the same as in Experiment 1. After having completed 10 practice trials with feedback, the experiment began. The task was a visual speeded classification in which the participants had to discriminate the size of the disc (large vs. small), while ignoring the simultaneously-presented task-irrelevant auditory pitches in an uninterrupted and rapid sequence (similar to

Gallace & Spence, 2006). The participants were invited to respond as rapidly as they could by pressing the left or right arrow on the keyboard, using the index and middle finger of their dominant hand, and this was established during training by considering an error any response provided 1500 msec or more after the stimulus onset. Half of the participants used the right arrow for the larger size and the left for the smaller one, while the other half did the opposite. The participants received no feedback about their responses during the experiment. Each trial started with the presentation of a fixation cross for a random timing between 1000 and 2000 msec followed by the presentation of a synchronized audiovisual pair of stimuli for 300 msec, with an inter-trial interval of 1500 msec during which time the participants had to respond.

The experiment was divided into three blocks of randomly administered trials. Each block was composed of two disc sizes (S/M, M/L or S/L). Every disc was presented with a simultaneous task-irrelevant pitch, chosen between six possible pitches for each block. Each disc could be paired with a congruent pitch – defined as a pitch higher (for smaller discs) or lower (for larger discs) than the preceding one – or with an incongruent one – defined as a pitch higher (for larger discs) or lower (for smaller discs) than the preceding one. The congruent sounds presented with each disc could be either the pitch that was preferred, on average, in Experiment 1, or by a pitch approximately one major third (4 semitones) higher or lower (see Fig. 4a). The incongruent sounds were chosen among the three used for the other disc in each block. Congruence and incongruence were obtained by manipulating the order of the trials, hence creating upwards or downwards pitch sequences: A congruent pairing would be an upwards interval coupled with a disc getting smaller than the one preceding it or a downwards interval coupled with a disc getting larger than the one preceding it (see Brunetti et al., 2018, for a similar procedure). For congruent pairings, this created a series of possible ascending and descending intervals between sounds that ranged from a minimum of 4 semitones – a major third (in S/M and M/L blocks) to a maximum of 32 semitones – two full

octaves and a sixth (in the S/L block; see Fig. 4b). This design allowed for a comparison between the effects of incongruent trials, congruent trials with average "preferred" sounds, and congruent trials with varying intervals.

Each block consisted of 126 trials, including 8 to 28 repetitions for incongruent sequences and for each of the different interval conditions for congruent sequences. The whole procedure lasted approximately 20 minutes.

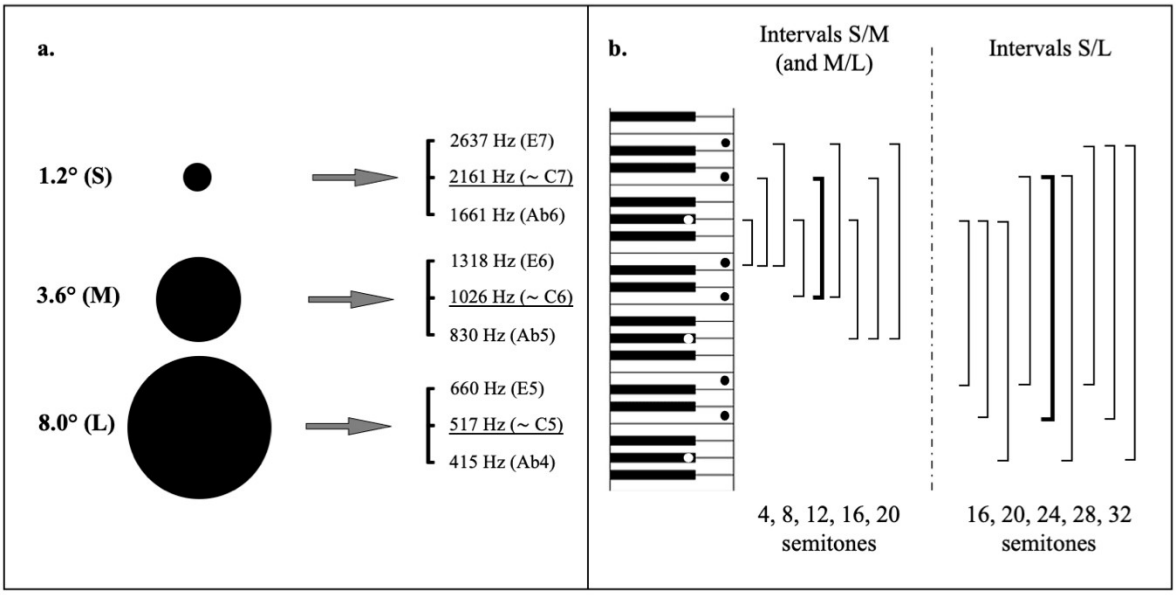


Figure 4. Panel (a) illustrates the specific pitches coupled with the three disc sizes used in Experiment 2. These pitches were the ones used for each size in the congruent trials. The middle underlined pitches are the average "preferred" ones according to Experiment 1 results. Panel (b) shows the intervals used in congruent trials in Experiment 2 (the keyboard is used only for illustration, as the pitches were all sinusoidal tones). Both blocks with Small and Medium and Medium and Large discs used intervals ranging from 4 to 20 semitones – the ones showed in the figure are the specific ones used in the S/M block, while those used in the M/L block were transposed according to the specific pitches used in that block. The S/L block featured larger intervals, ranging from 16 to 32 semitones in congruent trials. The intervals in bold brackets are those between "preferred" pitches taken from the results of Experiment 1.

Results

Initially, two repeated measures ANOVAs were conducted with the independent factor Pitch Interval, including 7 levels, and the dependent factor of Proportion Correct (PC). Since blocks S/M and M/L included the same intervals, this data was combined in one analysis (with the levels 4, 8, 12, 16, and 20 semitones; the 12 semitone intervals with "preferred" pitches and incongruent trials), while the data from the S/L block was analysed in a second ANOVA, with a similar design (with the levels 16, 20, 24, 28, and 32 semitones; the 20 semitones intervals with "preferred" pitches and incongruent trials). In case of the violation of the sphericity assumption, Greenhouse-Geisser corrected degrees of freedom were reported. Moreover, we carried out Bayesian repeated measure ANOVA to report the strength of evidence of each – main and post-hoc – effect (Keysers et al. 2020). The analysis of the S/M and M/L blocks revealed a significant effect of Pitch Interval: [F (3.53, 137.68) = 20.534, $p < .001$; $\eta^2_p = 0.345$, BF10 > 100]. A Holm-Bonferroni corrected post-hoc analysis showed that all PC scores for congruent intervals were significantly higher than the PC scores for incongruent trials ($p < .001$, and BF10> 100 in all comparisons; Fig. 5), highlighting a stable congruence effect. A significant difference was also revealed between intervals 4 vs. 20 semitones in congruent trials (p = .024, BF10 = 7.42). Moreover, Bayesian post-hoc analyses revealed evidence of differences between 4 vs. 8 semitones (BF10 = 7.43) and between 4 vs. 16 semitones (BF10 = 4.13. All of these findings indicate lower scores with the smaller pitch intervals for the congruent trials.

The analysis focusing on the S/L block also revealed a significant effect of Pitch Interval: [F (2.43, 94.68) = 19.3, $p < .001$, $\eta^2_p = 0.331$, BF10 > 100]. A Holm-Bonferroni corrected post-hoc analysis revealed the same congruence effect found in the S/M and M/L

blocks (all congruent interval trials were significantly different from incongruent trials; all *ps* < .001, all BF10>100; see Fig. 6).

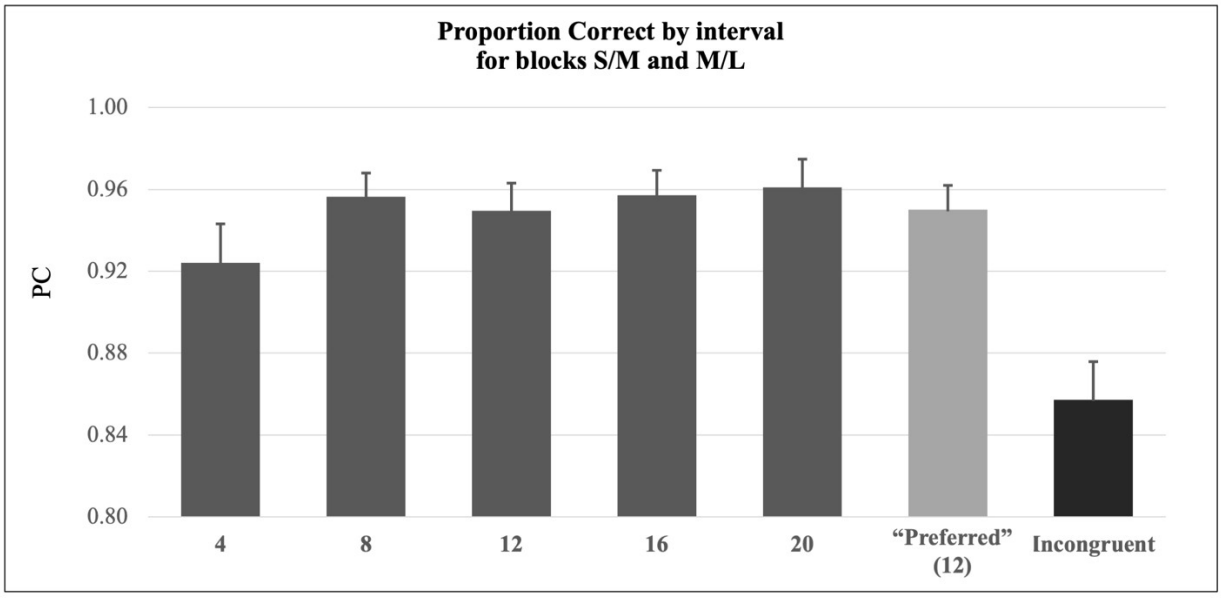


Figure 5. Speeded classification results from Experiment 2: S/M and M/L blocks. The proportion of correct (PC) responses shows a congruence effect that is practically identical for all congruent trials, regardless of the specific interval used, including the ones formed by the supposedly "Preferred" pitches. The error bars indicate the standard error of the means.

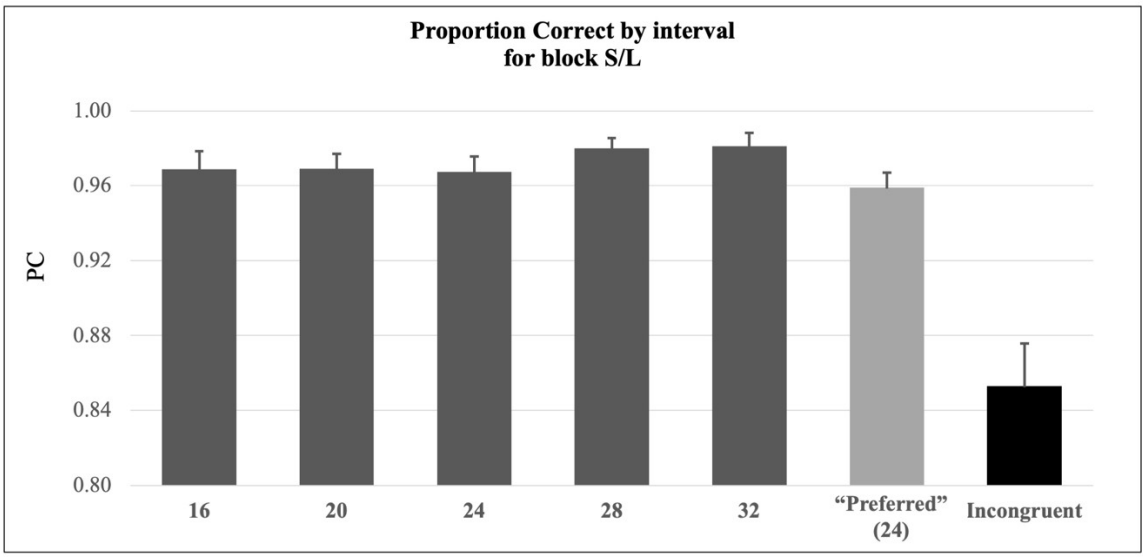


Figure 6. Speeded classification results from Experiment 2: S/L block. Similarly to the other blocks, a congruence effect for all congruent trials is evident. Once again, the effect highlight a polar correspondence, since it seems not dependent on the specific interval used, including the ones formed by the supposedly "Preferred" pitches. The error bars indicate the standard error of the means.

Second, Response Times (RT – only for correct responses) were analyzed as a dependent factor in two separate repeated measures ANOVAs, designed with the single independent factor Pitch Interval. Again, the data from the S/M and M/L blocks (with the levels 4, 8, 12, 16, and 20 semitones; the 12 semitones intervals with "preferred" pitches and incongruent trials) were combined, while the data from the S/L block was analysed in a second analysis, with a similar design (with the levels 16, 20, 24, 28, and 32 semitones; the 20 semitones intervals with "preferred" pitches and incongruent trials). The analysis for the S/M and M/L blocks revealed a significant effect of the Pitch Interval factor: [F (6, 234) = 18.095, $p < .001$; $\eta^2_p = 0.317$; BF10 > 100]. A Holm-Bonferroni corrected post-hoc analysis revealed that all RT scores for congruent intervals were significantly lower (faster) than the RT scores for incongruent trials (all $p < .04$, and BF10 > 3.40; Fig. 7). We also found a significant difference in congruent trials, specifically: 4-12 semitones ($p = .011$, BF10 =

42.6), 4-16 semitones (p < .001, BF10 >100), 4-20 semitones (p < .001, BF10 >100), 8-20 semitones (p =004, BF10 = 34.9), “Preferred” (12)-20 semitones (p = .011, BF10 = 25.1). Moreover, the Bayesian post-hoc showed a difference between 4-8 semitones (BF10 = 3.89), 4 - “Preferred” (12) semitones (BF = 3.39), and 12-20 semitones (BF10 = 3.93). All of these significant results were in the direction of lower RTs being associated with larger pitch intervals for the congruent trials.

Finally, the analysis focusing on the S/L block also revealed a significant effect of Pitch Interval: [F (6, 234) = 12.9, p < .001; η^2_p = 0.249, BF10 > 100]. Bonferroni corrected post-hoc analysis showed the same congruence effect found in the S/M and M/L blocks (all congruent interval trials were significantly faster than incongruent trials; all ps < .009, all BF10 > 16.9, Fig. 8). Moreover, a significant difference was also found between the congruent trials featuring “Preferred” (24) sounds and both 32 (p = .001, BF10 = 20.04; .05), and 20 semitones intervals (p = .026, BF10 = 5.62), and between 16 and 32 semitones intervals (p = .019, BF10 = 6.95). Moreover, the Bayesian post hoc analyses revealed also evidence of a difference between the congruent trials featuring “Preferred” (24) sounds and the 24 semitones intervals (BF10 = 8.91).

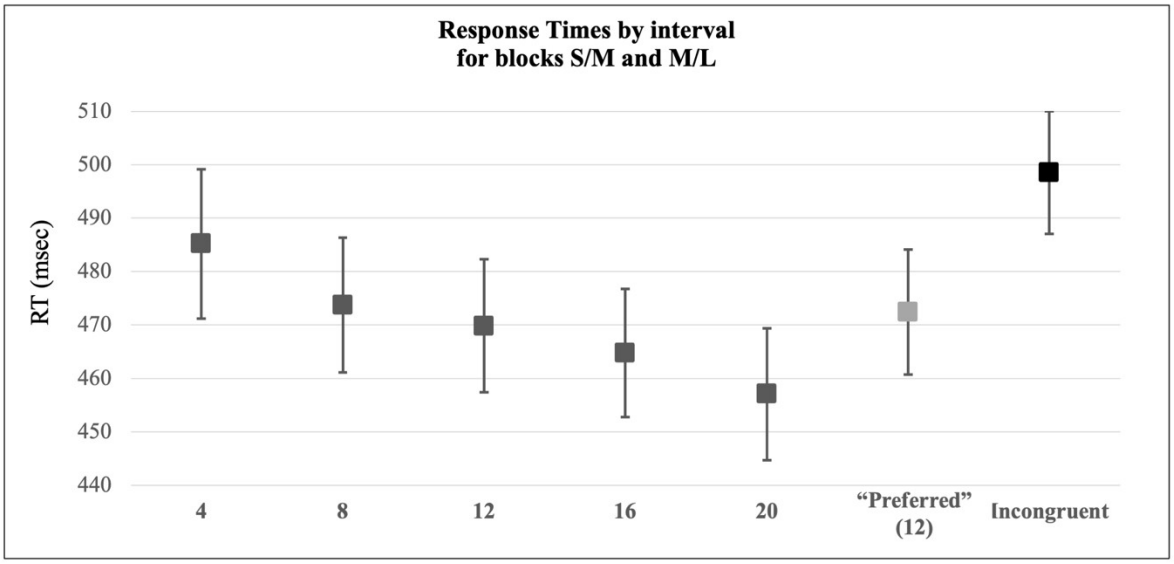


Figure 7. Speeded classification results from Experiment 2: S/M and M/L blocks. Response Times analysis shows a main congruence effect and evidence of faster RTs for larger intervals. Trials with supposedly "preferred" pitches do not appear to show any specific effect. The error bars indicate the standard error of the means.

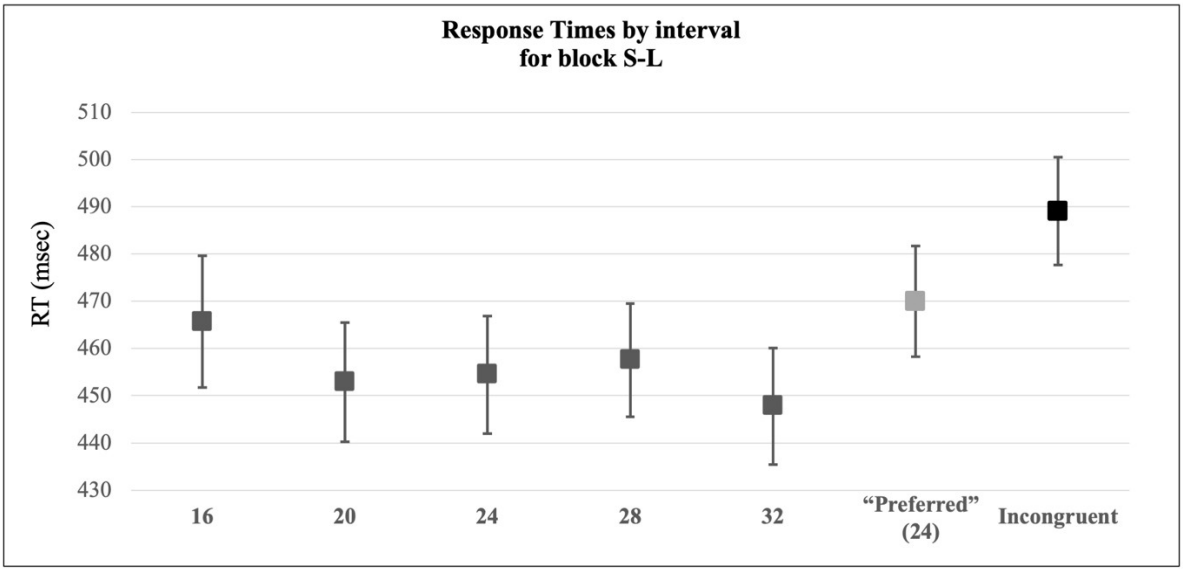


Figure 8. Speeded classification results from Experiment 2: S/L block. Response Times analysis shows a main congruence effect and some evidence of a larger interval effect. Trials with supposedly "preferred" pitches do not appear to show any specific effect. The error bars indicate the standard error of the means.

Discussion

The results of Experiment 2 confirmed a consistent and stable crossmodal congruence effect: Participants' performance was both slower (RT) and less accurate (PC) when the pitch sequence did not relatively match higher and lower pitches with larger and smaller discs, respectively, as in the incongruent trials (see Brunetti et al., 2018; Gallace & Spence, 2006; Spence, 2019a). However, Experiment 2 was designed to verify whether the pitches specifically matched to different sizes in Experiment 1 would enhance this well-known crossmodal congruence effect. This has not been the case: The intervals featuring average "preferred" pitches (as for the results of Exp. 1) did not affect the performance differently from other similar intervals. The intervals created using the average pitches matched by

participants in Exp. 1 showed a congruence effect that was seemingly identical to the same intervals not featuring “preferred” pitches in an independent sample of participants, taking part in Experiment 2. This lack of any specific enhancement for these pitches may be attributable to differences between the adjustment (Exp. 1) and the speeded categorization (Exp. 2) tasks. The main differences between these tasks concerns speed (i.e., the adjustment can be carried out without time constraints, while the speeded categorization task has to be performed rapidly), response type (i.e., the adjustment task asks for a subtle match between crossmodal stimuli while speeded categorization requires a rapid decision between two alternatives), and, crucially, attention. While in the adjustment procedure both modalities were task-relevant, the speeded categorization procedure requires the participant to focus on only one dimension/modality, while the other dimension/modality was task-irrelevant (in this case, participants were asked to attend to visual size, while auditory pitch had to be ignored).

In the light of these task differences, the discrepancy between the findings of Experiments 1 and 2 may be interpreted considering evidence that the size-pitch crossmodal correspondence seems to be regulated by top-down influences (Getz & Kubovy, 2018, p. 105). The difference in attentional engagement when considering the two tasks we proposed may be responsible of the different results we obtained: The “preferred” couplings emerged from Experiment 1 do not show any special effect in a task where attention is differently allocated. This would not only help to explain the difference in the results, but it could also explain why crossmodal correspondences sometimes seem to follow absolute couplings (see Guzman-Martinez et al., 2012) while mostly showing relative mappings (Brunetti et al., 2018; Spence, 2019a). This interpretation is also in line with Talsma and colleagues (2010), who proposed a framework whereby different task demands modulate the influence of top-down and bottom-up processes, and therefore attentional effects, in multisensory integration: changing the task, changes the involvement of attentional resources and thus the results.

Specifically, in an adjustment task, the need to pay attention to both dimensions, the deliberation involved in every trial, and possibly the lack of a time constraint, might favour those strategies based on careful evaluation and focusing on the response selection stage. Conversely, in a speeded categorization task, the focus on a single feature, the time pressure, and the simple evaluation, call for a completely different involvement of attentional resources (e.g., more based on the discrimination at the perceptual stage). The different engagement of attentional resources may be at the source of the apparently contradictory outcomes. Another possible explanation for these conflicting results may also be linked to the different groups of participants who were involved in Experiments 1 and 2, namely the inter-group variability could be linked to personal coupling preferences.

Nevertheless, Experiment 2 also reveals another aspect suggesting the dependence of such effects on perceptual bottom-up processes. The PC analyses did not reveal any specific interval effect, reminiscent of the constant effect found by Parise and Spence (2012) by using a different task and several different crossmodal correspondences. On the other hand, the RT analysis showed, mostly in S/M and M/L blocks, an interval effect: Congruent trials where more effective (e.g., generated faster responses) as a function of the pitch interval between consecutive trials. More specifically, the larger the interval, the faster the responses provided: When the contrast (e.g., the interval) between the pitches is perceptually more salient (e.g., larger), perception can benefit from the higher discriminability of the sounds, resulting in a faster decision. In Talsma et al.'s (2020) terms, more salient stimuli (e.g., more perceivable differences between pitches) may enhance performance by virtue of a perceptual bottom-up facilitation.

In summary, the results of Experiment 2 suggest two interesting conclusions: a top-down task-related influence (adjustment or speeding-up of categorization) and a bottom-up

influence related to stimulus salience (increased perceptual salience) on the strength of crossmodal pitch/size correspondences.

However, as pointed out by Spence (2019a, pp. 243-245), “the majority of researchers have tended to use only a relatively narrow range of frequencies”. To further test the hypothesis that salient stimuli (more perceivable differences between pitches) enhances pitch/size crossmodal correspondences, an additional experiment was designed. The aim of the Experiment 3 was to test whether the perceptual effect found in Experiment 2, connected to larger intervals, holds also with a wider range of pitches linked to only two disc sizes. Hence, Experiment 3 was carried out to further investigate if response times could be affected from congruent larger intervals, regardless of any specific pitch. Proportion of correct responses was expected to stay constant, consistently with the results of Experiment 2, showing only a congruency effect, without any modulation as a function of the width of the interval.

EXPERIMENT 3

Method

Participants

Sixty-seven undergraduates took part to the experiment for course credit (8 males; mean age = 20.7 years; SD = 2.96; range = 18–31 years). A post-hoc power analysis was carried out using G*Power 3.1.9.2 (F-tests, repeated-measures ANOVA). This analysis revealed that the current sample size guarantees a statistical power of more than 99% as computed by considering an effect size of 0.32, 7 measurements (corresponding to the 7 levels of the conducted ANOVA, see below the Results section), and a significance level of 0.05. All the participants reported normal or corrected-to-normal hearing and vision. The

participants were naïve as to the purpose of the study, which has been approved by the Ethics Review Board of the Università Europea di Roma.

Apparatus and materials

The apparatus was identical to that used in Experiment 2: a custom-made script in Max (v.8, Cycling'74) programming environment ran the experiment. The auditory component consisted of the same nine sinusoidal tones used in Experiment 2, with the same loudness adjustments. The visual stimuli consisted of 2 black discs of respectively 2.4° and 8.0° visual angle diameter; the discs were presented individually at the centre of a grey (50%) 21" display for 300 msec.

Procedure

The procedure consisted of a speeded classification task identical to that used in Experiment 2, with the same instructions, training, and response modality. This time, the three blocks all featured the same two discs: small (2.4°) and large (8.0°). Each block consisted of congruent and incongruent trials, each featuring upwards and downwards versions of the 8 possible intervals: 4, 8, 12, 16, 20, 24, 28, and 32 semitones. Each interval was repeated 40 times, balancing all conditions (320 trials in total). Each block consisted of 109 trials, giving rise to a total of 327 trials administered in a pseudo-random order (the additional 7 were dummy trials used for transitions and were not included in the subsequent data analyses). The repetitions included an equal number of congruent and incongruent trials, both upwards and downwards (e.g. congruent: from a trial with a larger disc to a trial with a smaller one with pitch getting higher and vice versa; incongruent: from a trial with a larger disc to a trial with a smaller one with pitch getting lower and vice versa). The whole procedure lasted for approximately 20 minutes.

Results

Four participants were removed from the analyses as their proportion of correct responses and/or RTs fell more than 2 SDs away from the sample average. Paired sample t-tests were performed to investigate whether the trial congruency affected the Proportion of Correct responses (PC) and the Response Time. We also carried out Bayesian paired t-test to report the strength of evidence of each effect (Keysers et al. 2020). Results were significant for both analyses: PC[t(62) = 2.97, p = .004, Cohen's d = 0.374, BF10 = 7.32]] (Congruent trials' mean = .943, SE = .007; Incongruent trails' mean = .932, SE = .007); RT - [t(62) = 4.27, p < .001, Cohen's d = 0.538, BF10 > 100] (Congruent trials' mean = 471 msec, SE = 6.36; Incongruent trails' mean = 476 msec, SE = 6.83).

The analysis followed through to investigate the influence of the intervals used in the experiment. Intervals 4 and 8, 12 and 16, 20 and 24, 28 and 32 were grouped; then the discrepancy (Delta) between Incongruent and Congruent trials for each group was calculated, to quantify the advantage of congruent trials in PC and RT. Each Delta was obtained by subtracting the average score (PC or RT) of congruent trials from the average score of incongruent trials to obtain a measure expressing the amount of facilitation or inhibition of performance: if positive, the Delta indicates higher scores in incongruent trials, while if negative it indicates higher scores in congruent trials.

Two repeated measures ANOVAs were carried out, featuring Interval as a single factor with 4 levels: small (4-8 semitones), medium (12-16 semitones), large (20-24 semitones), and very large (28-32 semitones). Greenhouse-Geisser corrected degrees of freedom were reported in the case of sphericity assumption violation. Moreover, Bayesian repeated measure ANOVA was performed to report the strength of evidence of each - main and post-hoc – effect (Keysers et al. 2020). The first ANOVA was carried out with Delta PC

as a dependent factor and did not yield significant results: [F (3, 186) = 2.19, p = .09; η^2_p = 0.034, BF10 = 0.36]. The second ANOVA was carried out using Delta RT as a dependent factor. The analysis revealed significant results: [F (3, 186) = 5.12, p = .002; η^2_p = 0.076, BF10 =18.84] (see Table 1 for the average scores obtained with both dependent factors). RT discrepancies between congruent and incongruent trials significantly increase as the interval between task-irrelevant pitches increases (see Fig. 9). Post-hoc Holm-Bonferroni tests revealed a significant difference between the intervals small (4-8 semitones) and very large (28-32 semitones; p = .001), while none of the other comparisons reached significance. Moreover, Bayesian post-hoc analyses confirmed the decisive evidence for this difference (4-8 vs. 28-32, BF10 > 100).

Experiment 3 Delta scores

Interval	**Delta PC (not sign.)**	**Delta RT (p < .01)**
Small (4-8 st)	-.013 (.058)	-.917 (16.655)
Medium (12-16 st)	-.002 (.052)	3.932 (16.271)
Large (20-24 st)	-.026 (.060)	6.413 (18.164)
Very large (28-32 st)	-.005 (.058)	10.805 (19.103)

Table 1. Speeded classification results from Experiment 3. Analysis of Delta PC scores did not yield significant results, while the analysis of Delta RT scores was significant. Values are reported in the format Mean (SD).

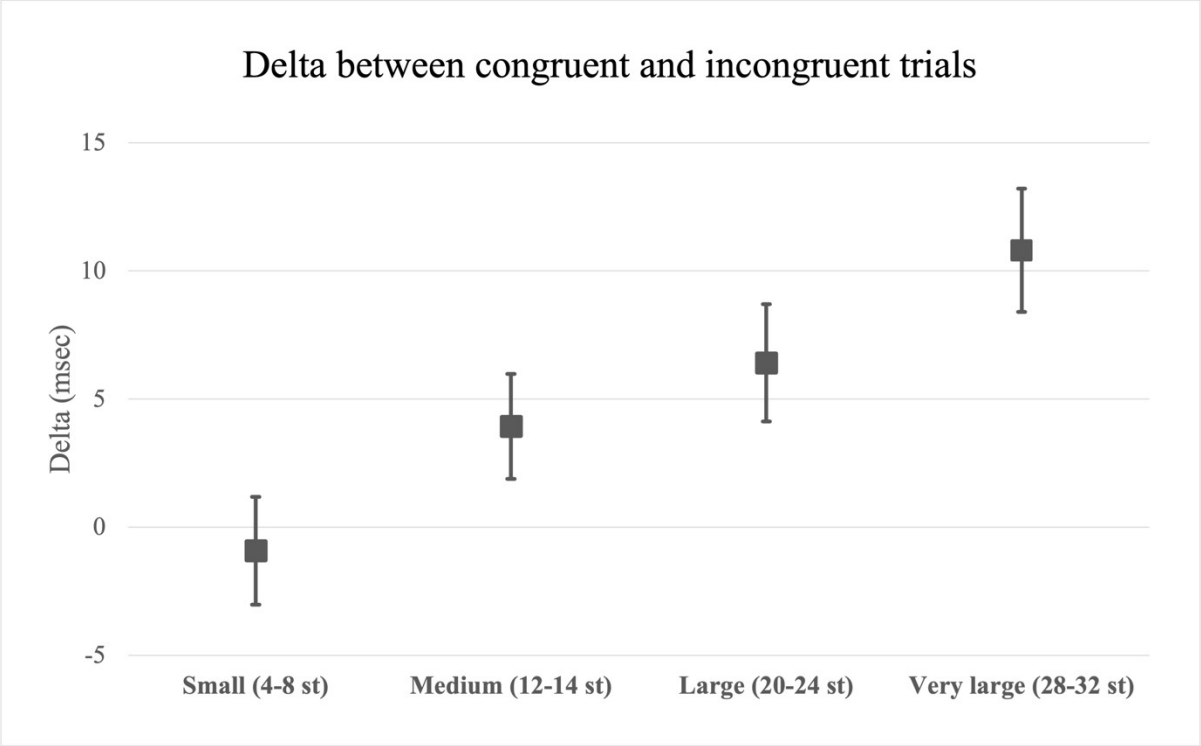


Figure 9. Speeded classification results from Experiment 3. Delta RT are obtained by subtracting the average RT of congruent trials from the average RT of incongruent trials: Positive scores indicate slower responses to incongruent trials. Analyses showed a significant effect of interval. The error bars indicate the standard error of the means.

Discussion

The results of Experiment 3 confirmed the perceptual interval effect suggested by the results of Experiment 2. Exploring the effects of different intervals systematically revealed that crossmodal size-pitch congruence (high pitch/small size, low pitch/large size) yields greater effects with larger intervals. Experiment 3 explored a rather large variety of intervals, ranging from a major third up to almost 3 octaves, in musical terms. These results suggest that the size-pitch congruence is not constrained by specific ranges and that the effect is enhanced by a greater perceptual difference between pitches: In this sense, there would

appear to be no boundaries, at least within the range of stimuli tested in this study, to the matching between the different sensorial dimensions involved. This finding confirms and expands the results of Brunetti and colleagues (2018): While they showed how pairings could adapt to different pitch contexts, Experiment 3 demonstrated that the effects are maximized when the pitches are the farthest apart from each other. Notably, this effect is only evident in the RT data, while, as hypothesized, it does not systematically affect the PC data. Consistently, crossmodal correspondence research often mainly records effects on RT data, with effects on accuracy varying across studies (e.g., Evans, 2020; Evans & Treisman, 2010).

The larger intervals may show a stronger congruence effect because larger differences in pitch are perceptually more salient than smaller ones, outside of a clear melodic tonal context (within a melodic tonal context, salience rules are different and do not follow magnitude alone, see Trainor & Trehub, 1992). As noted in the discussion of Experiment 2, this appears to agree with the idea that the crossmodal congruence effect in speeded categorization can benefit from a salience-driven enhancement, thus corroborating the idea that different tasks (and correspondences) may be affected differently by top-down and bottom-up influences (see Evans, 2020; Talsma et al., 2010).

General Discussion

All the experiments taken together offer insightful new evidence in the study of the size-pitch correspondence and possibly on the study of multisensory integration, more generally. The experiments reported here explored the extent of possible congruent mappings using the method of adjustment and visual speeded categorization, with a relatively wide range of visual and auditory stimuli (cf. Spence, 2019a): Experiment 1 seems to demonstrate that pitch/size crossmodal correspondences are due to quasi-absolute mappings related to bottom-up processes; Experiment 2 revealed evidence of an attentional top-down effect

related to the task used in the paradigm, and a perceptual bottom-up effect related to the interval distance of pitches; Experiment 3 supports the enhanced bottom-up effect due to a larger perceptual salience.

Specifically, the results of Experiment 1 reveal that, when paying attention to both sensory dimensions at the same time with the task of pairing them, a consistent relationship appears among audiovisual intervals. These findings extend the previous literature exploring other crossmodal correspondences with the method of adjustment (Guzman-Martinez *et al.*, 2012; Orchard-Mills *et al.*, 2013) using a larger sample of participants, thus providing a better estimation of the effects. As in other studies, the monotonic relationship was found to link the two dimensions with some inconsistencies across different conditions. While Orchard-Mills et al. (2013) documented that the starting speed of the amplitude modulation significantly affected the pairings, here a task effect was found, highlighting a difference between the case where the participants were asked to adjust pitch based on size and the opposite task (e.g., adjusting size based on pitch). The results of the two tasks do not overlap, showing an effect due to task demands: an asymmetry due either to where attention was primarily put (e.g., on adjusting Pitch in response to a provided Size or the opposite) or to the specific stimuli set or range used in the two tasks. This kind of asymmetry has been reported previously in research on the crossmodal correspondences (Spence, 2019b) and a potential interpretation involves the role of attention in the adjustment task, reflecting top-down modulations (related to where attention is focused and to the specific sensory range available in the adjustment tasks).

The results of Experiment 2 corroborate this notion. Experiment 2 explored the possibility whether some “preferred” pairings (selected on the basis of Experiment 1) may modulate the impact of size-pitch correspondence in a speeded categorization task, whereas only one dimension was task-relevant (i.e., vision). The design of Experiment 2 included

these pairings estimated through Experiment 1 in a large set of stimuli, ranging across 3 visual sizes (i.e., the attended dimension) and 6 pitch intervals at various pitch heights (i.e., the task-irrelevant dimension). The results clearly revealed that the supposedly “preferred” pairings did not actually affect the performance in any way, affecting the performance indistinguishably from similar size intervals. In fact, Experiment 2 hinted at another relationship, this time clearly linear and dependent on the difference between the low and high pitches. The large stimulus set allowed us to observe some signs of an effect of the larger pitch intervals used in the pairings: namely, when the low and high pitches were more distant from each other, the effect of the task-irrelevant dimension was stronger, resulting in a larger difference between congruent and incongruent trials. This effect, while not pinpointed in any other study, may well be indirectly confirmed by the pitch intervals usually used in other pitch-related correspondence studies. As Spence’s (2019a) review makes clear, while effects can be obtained with small intervals (e.g., 600-680 Hz), the majority of studies have used pitches that are separated by quite a large interval (typically 300-4500 Hz; Spence, 2019a, pp. 243-244). This, along with the results presented here, and especially the fact that the supposedly “preferred” pairings did not yield any enhanced effects in Experiment 2, appears ultimately to confirm the relative (as opposed to absolute) nature of the size-pitch correspondence (Brunetti et al., 2018; Spence, 2019a). These results also extend the available literature by showing that size-pitch correspondences are modulated by a strategic/top-down influence. An alternative explanation for the absence of an enhanced effect of the "preferred" pairings could be related to the fact that the preferred pairings are based on the averages of all the participants who took part in Experiment 1, and may not necessarily be consistent with the size-pitch crossmodal mappings of each participant. In future studies, it might be interesting to repeat the experiment with a procedure that varies according to each participant’s "preferred" pairing.

Finally, the results of Experiment 3 demonstrate that sensory contrast (e.g., larger pitch intervals) can enhance the congruence effect, thus showing a bottom-up influence on the correspondence. In Experiment 3, 2 visual sizes and a set of 7 pitch intervals were used. This approach showed larger congruence effects in terms of RTs compared to Experiment 2, suggesting that the size-pitch correspondence has virtually no set boundaries. This clearly reveals a bottom-up influence, linked to perceptual salience, that can significantly affect the overall correspondence congruence effect. Overall, the results of Experiments 1-3 reveal that size-pitch correspondences are sensitive to task-related top-down factors (different tasks require different attentional engagement) as well as stimulus-related bottom-up influences (related to perceptual salience).

This evidence falls well within the framework proposed by Talsma et al. (2010), specifically when they state that salient stimuli tend to facilitate multisensory integration processes even in a context where top-down attention is the main responsible for the task at hand (p. 402).

Finally, in the specific context of pitch-based correspondences, the interplay between top-down and bottom-up processes is highlighted from different studies that have demonstrated that musically trained participants show significantly stronger associations than untrained ones (Küssner & Leech-Wilkinson, 2014). However, this top-down effect influences the degree of crossmodal mapping, rather than the direction or the type of mapping that, instead, seems to depend on stimuli (Eitan & Granot, 2006; Eitan & Timmers, 2010; Küssner & Leech-Wilkinson, 2014). To further clarify the interplay between top-down and bottom-up influences, recruiting participants with musical training or absolute pitch (Takeuchi & Hulse, 1993) could offer valuable insight into the mechanisms behind the size-pitch congruence effects (i.e., as in the recent contribution by Ihalainen, Kotsaridis, Vivas, & Paraskevopoulos, 2023). Investigating how individuals with experience (in the case of

musicians), or those with extraordinary memory and recognition skills (in the case of those individuals with absolute pitch), would offer additional information concerning top-down influences (related to skill) and bottom-up effects (stimulus-driven). More specifically, investigating whether participants familiar with pitch organization would react differently to pitch intervals (Aruffo, Goldstone, & Earn, 2014); or whether participants able to clearly remember and recognize absolute pitch heights show relative mappings (Dooley & Deutsch, 2011), would help to shed new light on the influence of top-down and bottom-up processes.

Overall, the results presented here confirm the size-pitch correspondence as modulated by tasks and completely relative in nature, therefore potentially hinged in top-down attentional processes (Getz & Kubovy, 2018; Klapetek et al., 2012); however, they also demonstrate that its effects can be significantly modulated by bottom-up influence, in this case obtained by larger pitch intervals. This study calls for further investigation about all crossmodal correspondences, to understand if and how much they could be differentially influenced by top-down and bottom-up processes. Some previous studies tend to differentiate correspondences in terms of being of a top-down or bottom-up nature perhaps too sharply (Getz & Kubovy, 2018), while maybe the interplay between influences coming from the stimuli and from the cognitive system should be explored always starting from the perspective of a compresence of all factors. In this sense, each correspondence could reveal a differential dependence on top-down and bottom-up influences that could help shedding light on the nature of each one, offering another key element to differentiate between them (Spence & Deroy, 2013) and further reveal their nature.

Authors' Contributions

RB, CDG, and AI designed the experiments; AI and CDG organized the population recruitment and performed the experiments; VS and CS offered valuable insights on the

literature; TP processed and performed the behavioural data analysis; RB, AI, and CDG wrote the manuscript. All authors revised and approved the manuscript.

Data availability: All data have been made publicly available at the Research Box and can be accessed at https://researchbox.org/1933&PEER_REVIEW_passcode=SAMEPX. None of the experiments was preregistered.

References

Alais, D. & Burr, D. (2019). Cue combination within a Bayesian framework. In Lee, K. C., Wallace, M. T., Coffin, A. B., Popper A. N., Fay, R. R. (Eds), *Multisensory Processes. The Auditory Perspective* (pp. 9-31). Cham, CH: Springer International. DOI 10.1007/978-3-030-10461-0_2.

Alsius, A., Navarra, J., Campbell, R., & Soto-Faraco, S. (2005). Audiovisual integration of speech falters under high attention demands. *Current Biology*, *15*(9), 839-843. DOI 10.1016/j.cub.2005.03.046

Andersen, T. S., Tiippana, K., & Sams, M. (2005). Maximum likelihood integration of rapid flashes and beeps. *Neuroscience Letters*, *380*(1-2), 155-160. DOI 10.1016/j.neulet.2005.01.030

Aruffo, C., Goldstone, R. L., & Earn, D. J. (2014). Absolute judgment of musical interval width. *Music Perception: An Interdisciplinary Journal*, 32(2), 186-200. DOI 10.1525/mp.2014.32.2.186

Ben-Artzi, E., & Marks, L. E. (1995). Visual-auditory interaction in speeded classification: Role of stimulus difference. *Perception & Psychophysics, 57,* 1151-1162. DOI 10.3758/BF03208371

Benoit, K. (2011). Linear regression models with logarithmic transformations. *London School of Economics, London, 22*(1), 23-36.

Brunetti, R., Indraccolo, A., Del Gatto, C., Spence, C., & Santangelo, V. (2018). Are crossmodal correspondences relative or absolute? Sequential effects on speeded classification. *Attention, Perception, & Psychophysics, 80*(2), 527-534. DOI 10.3758/s13414-017-1445-z

Brunetti, R., Indraccolo, A., Mastroberardino, S., Spence, C., & Santangelo, V. (2017). The impact of crossmodal correspondences on working memory performance. *Journal of Experimental Psychology: Human Perception and Performance, 43,* 819-831. DOI 10.1037/xhp0000348

Cardozo, B. L. (1965). Adjusting the method of adjustment: SD vs DL. *Journal of the Acoustical Society of America*, *37*(5), 786-792. DOI 10.1121/1.1909439

Chiou, R., & Rich, A. N. (2012). Cross-modality correspondence between pitch and spatial location modulates attentional orienting. *Perception, 41,* 339-353. DOI 10.1068/p7161

Chiou, R., & Rich, A. N. (2015). Volitional mechanisms mediate the cuing effect of pitch on attention orienting: The influences of perceptual difficulty and response pressure. *Perception, 44*, 169-182. DOI 10.1068/p7699

Del Gatto, C., Brunetti, R., & Delogu, F. (2016). Cross-modal and intra-modal binding between identity and location in spatial WM: The identity of objects does not help recalling their locations. *Memory, 24*, 603-615. DOI 0.1080/09658211.2015.1034137

Deroy, O., Crisinel, A.-S., & Spence, C. (2013). Crossmodal correspondences between odors and contingent features: Odors, musical notes, and geometrical shapes. *Psychonomic Bulletin and Review, 20*, 878-896. DOI 10.3758/s13423-013-0397-0

Deroy, O., & Spence, C. (2013). Why we are not all synesthetes (not even weakly so). *Psychonomic Bulletin & Review*, *20*(4), 643-664. DOI 10.3758/s13423-013-0387-2

Dooley, K., & Deutsch, D. (2011). Absolute pitch is correlated with high performance on relative pitch tasks. *Journal of the Acoustical Society of America*, 129(4_Supplement), 2582. DOI 10.1121/1.3588532

Duyar, A., Pavan, A., & Kafaligonul, H. (2022). Attentional modulations of audiovisual interactions in apparent motion: Temporal ventriloquism effects on perceived visual speed. *Attention, Perception & Psychophysics, 84*(7), 2167-2185. DOI 10.3758/s13414-022-02555-7

Eitan, Z., & Granot, R. Y. (2006). How music moves: Musical parameters and listeners images of motion. *Music Perception*, *23*(3), 221-248. DOI 10.1525/mp.2006.23.3.221

Eitan, Z., & Timmers, R. (2010). Beethoven's last piano sonata and those who follow crocodiles: Cross-domain mappings of auditory pitch in a musical context. *Cognition*, *114*(3), 405-422. DOI 10.1016/j.cognition.2009.10.013

Erlebacher, A., & Sekuler, R. (1971). Response frequency equalization: A bias model for psychophysics. *Perception & Psychophysics, 9*, 315-320. DOI 10.3758/BF03212657

Evans, K. K. (2020). The role of selective attention in cross-modal interactions between auditory and visual features. *Cognition, 196*:104119. DOI 10.1016/j.cognition.2019.104119

Evans, K. K., & Treisman, A. (2010). Natural cross-modal mappings between visual and auditory features. *Journal of Vision, 10*, 1-12. DOI 10.1167/10.1.6

Gallace, A., & Spence, C. (2006). Multisensory synesthetic interactions in the speeded classification of visual size. *Perception & Psychophysics, 68*(7), 1191-1203. DOI 10.3758/bf03193720

Getz, L. M., & Kubovy, M. (2018). Questioning the automaticity of audiovisual correspondences. *Cognition, 175*, 101-108. DOI 10.1016/j.cognition.2018.02.015

Guzman-Martinez, E., Ortega, L., Grabowecky, M., Mossbridge, J., & Suzuki, S. (2012). Interactive coding of visual spatial frequency and auditory amplitude-modulation rate. *Current Biology, 22*, 383-388. DOI 10.1016/j.cub.2012.01.004

Hutmacher, F. (2019). Why is there so much more research on vision than on any other sensory modality? *Frontiers in Psychology*, *10*:2246. DOI 10.3389/fpsyg.2019.02246.

Ihalainen, R., Kotsaridis, G., Vivas, A.B., & Paraskevopoulos, E. (2023). The relationship between musical training and the processing of audiovisual correspondences: Evidence from a reaction time task. *PLoS ONE 18*(4): e0282691. DOI 10.1371/journal.pone.0282691

Jamal, Y., Lacey, S., Nygaard, L., & Sathian, K. (2017). Interactions between auditory elevation, auditory pitch and visual elevation during multisensory perception. *Multisensory Research*, *30*(3-5), 287-306. DOI 10.1163/22134808-00002553

Kalan, A. K., Mundry, R., & Boesch, C. (2015). Wild chimpanzees modify food call structure with respect to tree size for a particular fruit species. *Animal Behaviour, 101*, 1-9. DOI 10.1016/j.anbehav.2014.12.011

Keysers, C., Gazzola, V., & Wagenmakers, E.J. (2020). Using Bayes factor hypothesis testing in neuroscience to establish evidence of absence. *Nature Neuroscience, 23*, 788–799. DOI 10.1038/s41593-020-0660-4

Klapetek, A., Ngo, M. K., & Spence, C. (2012). Does crossmodal correspondence modulate the facilitatory effect of auditory cues on visual search? *Attention, Perception, & Psychophysics, 74*, 1154-1167. DOI 10.3758/s13414-012-0317-9

Küssner, M. B., & Leech-Wilkinson, D. (2014). Investigating the influence of musical training on cross-modal correspondences and sensorimotor skills in a real-time drawing paradigm. *Psychology of Music*, *42*(3), 448-469. DOI 10.1177/0305735613482022

Loconsole, M., Stancher, G., & Versace, E. (2023). Crossmodal association between visual and acoustic cues in a tortoise (*Testudo hermanni*). *Biology Letters, 19*(7). DOI 10.1098/rsbl.2023.0265

Marks, L. E. (2004). Cross-modal interactions in speeded classification. In G. A. Calvert, C. Spence, & B. E. Stein (Eds.), *Handbook of Multisensory Processes* (pp. 85-105). Cambridge, MA: MIT Press. DOI https://doi.org/10.7551/mitpress/3422.003.0009

Martino, G., & Marks, L. E. (2001). Synesthesia: Strong and weak. *Current Directions in Psychological Science, 10,* 61-65. DOI 10.1111/1467-8721.00116

McCormick, K., Lacey, S., Stilla, R., Nygaard, L. C., & Sathian, K. (2018). Neural basis of the crossmodal correspondence between auditory pitch and visuospatial elevation. *Neuropsychologia, 112*, 19-30. DOI 10.1016/j.neuropsychologia.2018.02.029

Melara, R. D. (1989). Dimensional interaction between color and pitch. *Journal of Experimental Psychology: Human Perception and Performance, 15,* 69-79. DOI 10.1037/0096-1523.15.1.69

Melara, R. D., & O'Brien, T. P. (1987). Interaction between synesthetically corresponding dimensions. *Journal of Experimental Psychology: General, 116,* 323-336. DOI 10.1037/0096-3445.116.4.323

Mellers, B., & Birnbaum, M. H. (1982). Loci of contextual effects in judgment. *Journal of Experimental Psychology: Human Perception & Performance, 8*(4), 582-601. DOI 10.1037/0096-1523.8.4.582

O'Boyle, M. W., & Tarte, R. D. (1980). Implications for phonetic symbolism: The relationship between pure tones and geometric figures. *Journal of Psycholinguistic Research, 9*, 535-544. DOI 10.1007/BF01068115

Orchard-Mills, E., Van der Burg, E., & Alais, D. (2013). Amplitude-modulated auditory stimuli influence selection of visual spatial frequencies. *Journal of Vision, 13,* 1-17. DOI 10.1167/13.3.6

Parise, C. V., Knorre, K., & Ernst, M. O. (2014). Natural auditory scene statistics shapes human spatial hearing. *Proceedings of the National Academy of Sciences of the USA, 111*, 6104-6108. DOI 10.1073/pnas.1322705111

Parise, C., & Spence, C. (2009). 'When birds of a feather flock together': Synesthetic correspondences modulate audiovisual integration in non-synesthetes. *PLoS ONE, 4*(5): e5664. DOI 10.1371/journal.pone.0005664.

Parise, C. V. & Spence, C. (2012). Audiovisual crossmodal correspondences and sound symbolism: An IAT study. *Experimental Brain Research, 220,* 319-333. DOI 10.1371/journal.pone.0005664

Posner, M. I., Nissen, M. J., & Klein, R. M. (1976). Visual dominance: An information-processing account of its origins and significance. *Psychological Review, 83,* 157-171. DOI 10.1037/0033-295X.83.2.157

Santangelo, V., & Macaluso, E. (2012). Spatial attention and audiovisual processing. In B. E. Stein (Ed.), *The New Handbook of Multisensory Processing* (pp. 359-370). Cambridge, MA: MIT Press. DOI 10.7551/mitpress/8466.003.0031

Sciortino, P., & Kayser, C. (2023). Steady state visual evoked potentials reveal a signature of the pitch-size crossmodal association in visual cortex. *NeuroImage*, *273*: 120093. DOI 10.1016/j.neuroimage.2023.120093

Sekuler, R., & Erlebacher, A. (1971). The invalidity of "invalid results from the method of constant stimuli": A common artifact in the methods of psychophysics. *Perception & Psychophysics, 9,* 309-311. DOI 10.3758/BF03212655

Spence, C. (2011). Crossmodal correspondences: A tutorial review. *Attention, Perception, & Psychophysics, 73,* 971-995. DOI 10.3758/s13414-010-0073-7

Spence, C. (2019a). On the relative nature of (pitch-based) crossmodal correspondences. *Multisensory Research, 32,* 235-265. DOI 10.1163/22134808-20191407

Spence, C. (2019b). On the relationship(s) between color and taste/flavor. *Experimental Psychology, 66*(2), 99-111. DOI 10.1027/1618-3169/a000439

Spence, C., & Deroy, O. (2013). How automatic are crossmodal correspondences? *Consciousness and Cognition, 22*(1), 245-260. DOI 10.1016/j.concog.2012.12.006

Spence, C., & Di Stefano, N. (2022). Coloured hearing, colour music, colour organs, and the search for perceptually meaningful correspondences between colour and pitch. *i-Perception, 13*(3):1-42. DOI 10.1177/20416695221092802

Stevens, S. S. (1957). On the psychophysical law. *Psychological Review, 64,* 153-181.

Talsma, D., Senkowski, D., Soto-Faraco, S., & Woldorff, M. G. (2010). The multifaceted interplay between attention and multisensory integration. *Trends in Cognitive Sciences, 14*, 400-410. DOI 10.1016/j.tics.2010.06.008

Taylor, A. M., Reby, D., & McComb, K. (2011). Cross modal perception of body size in domestic dogs *(Canis familiaris). PLoS ONE, 6*: e17069. DOI 10.1371/journal.pone.0017069

Takeuchi, A. H., & Hulse, S. H. (1993). Absolute pitch. *Psychological Bulletin*, *113*(2), 345-361. DOI 10.1037/0033-2909.113.2.345

Trainor, L. J., & Trehub, S. E. (1992). A comparison of infants' and adults' sensitivity to western musical structure. *Journal of Experimental Psychology: Human Perception and Performance, 18*(2), 394-402. DOI 10.1037/0096-1523.18.2.394

Utts, J. M. (1982). The rainbow test for lack of fit in regression. *Communications in Statistics-Theory and Methods, 11*(24), 2801-2815. DOI 10.1080/03610928208828423

Varghese, L., Mathias, S. R., Bensussen, S., Chou, K., Goldberg, H. R., Sun, Y., Sekuler, R., & Shinn-Cunningham, B. G. (2017). Bi-directional audiovisual influences on temporal modulation discrimination. *The Journal of the Acoustical Society of America*, *141*(4), 2474-2488. DOI 10.1121/1.4979470

Walker, P. (2016). Cross-sensory correspondences: A theoretical framework and their relevance to music. *Psychomusicology: Music, Mind, and Brain, 26*, 103-116. DOI 10.1037/pmu0000130

Welch, R. B., & Warren, D. H. (1980). Immediate perceptual response to intersensory discrepancy. *Psychological Bulletin, 88*, 638-667. DOI 10.1037/0033-2909.88.3.638